OEG Publication

Aguado de Cea G, Álvarez de Mon I, Gómez-Pérez A, Pareja-Lora A

OntoTag: XML/RDF(S)/OWL Semantic Web Page Annotation in ContentWeb

Proceedings of the 3rd Workshop on NLP and XML – Language Technology and the Semantic Web (NLPXML-2003).
10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03).
Edited by N. Ide. Association for Computational Linguistics. April 2003.
Budapest, Hungry.
Pages: 25 to 30

OntoTag: XML / RDF(S) / OWL Semantic Web Page Annotation in ContentWeb

Guadalupe Aguado de Cea
DLACT / LIA-PLN, UPM
lupe@fi.upm.es

Inmaculada Álvarez-de-Mon
DLACT / LIA-PLN, UPM
ialvarez@euitt.upm.es

Asunción Gómez-Pérez
DIA / LIA, UPM
asun@fi.upm.es

Antonio Pareja-Lora
DSIP, UCM / LIA-PLN, UPM
apareja@sip.ucm.es

Abstract

As the Semantic Web and the Language Resources research fields become closer, the need for standard representation formats and languages gets clearer, specially taking into account the increasing need for cooperation and interoperability between both fields that is being set forth. The purpose of this paper is to present how this process of standardisation and integration is being achieved in *ContentWeb* by means of *OntoTag*, a multi-level (also multi-purpose and possibly multi-language) hybrid (ontologic and linguistic) platform for Semantic Web annotation, designed according to EAGLES standards and implemented with last generation Semantic Web languages (XML/ RDF(S)/OWL).

1 Introduction

The main pillar sustaining the development of what we understand by *Semantic Web* –"the conceptual structuring of the web in an explicit machine-readable way" (Berners-Lee et al., 1999)– is enabling computers to understand the meaning (the semantics) of written texts and web pages by making meaning explicit, in one way or another, to computers. Even though the automatic process of information is being eased and intensive research is being carried out in this area, still relevant information-related tasks such as information access, extraction and interpretation cannot be wholly performed by computers. In this context, *semantic annotation*, since it makes meaning explicit, has become a key topic.

Thus, much research is being done on semantic annotation, not only in the field of *Corpus Linguistics* (Wilson & Thomas, 1997; Schmidt, 1988), but also in the field of *Artificial Intelligence*, following the guidelines of the *Semantic Web* initiative (Benjamins et al., 1999), (Motta et al., 1999), (Luke et al., 2000), (Staab et al., 2000): great efforts are being devoted to the design and application of models and formalisms for the semantic annotation of web pages in order to make these documents more machine-readable. Far from being irreconcilable, the kind of annotations developed in these two fields can be considered complementary and mutually enlightening (Aguado–de Cea et al., 2002a). However, the need for a common vocabulary, as well as resource availability, interoperability and sharing, to join the efforts of both communities' researchers is still only partially (if at all) fulfilled. It is in this point where standardisation can play a key role in the development, expansion and success of the Semantic Web.

From this point of view, the benefits of Artificial Intelligence and Corpus Linguistics joint work will be invaluable. Both fields together can provide the Semantic Web with a suitable hybrid annotation, incorporating the main linguistic levels from Corpus Linguistics (since, as shown in Aguado–de Cea et al. (2002a)

and in Buitelaar et al. (2003), meaning cannot be considered confined to the semantic level of linguistic description) as well as the kind of (ontological) information that Artificial Intelligence annotations include and even broadening the scope of both of them to bear new fashions and schemes of corpus (semantic) annotation.

This paper shows the preliminary research results of the project *ContentWeb* on how Corpus Linguistics at all its main levels and Artificial Intelligence annotations can be joined together in a hybrid standardising annotation scheme by means of the platform *OntoTag* (which makes extensive use of RDF(S), XML and OWL for its annotations). It is organised as follows: firstly, a very brief state of the art in Corpus Linguistics' recommendations, criteria and guidelines for annotation will be presented –section 2–. Secondly, in section 3, the aforementioned *ContentWeb* project is introduced, and its platform for hybrid annotation, *OntoTag*, is described in section 4. OntoTag's ontologies for linguistic standard compliance are presented in section 5. Then, some conclusions will be outlined –section 6–, followed by the acknowledgments section and, finally, the references.

2 Corpus Linguistics Annotation – towards Standardisation.

Even though much research has been carried out by ontologists in the Semantic Web field on the semantic annotation of web pages (Aguado–de Cea et al., 2002b; EsperOnto, 2003) it is in the field of Corpus Linguistics where most standards, criteria and recommendations on annotation can be found. In EAGLES (1996a), a list of the main different levels of linguistic annotation can be found, namely: lemma, morphosyntactic, syntactic, semantic and discourse annotation. They are shown in Figure 1 (Annotation Level Pyramid), together with their corresponding tools (Linguistic Tool Stack) and applicable criteria, recommendations and guidelines (Linguistic Annotation Criteria Heap). A deep analysis of these concepts and their potential value for the Semantic Web can be found in EsperOnto (2003) and Aguado–de Cea et al. (2002a).

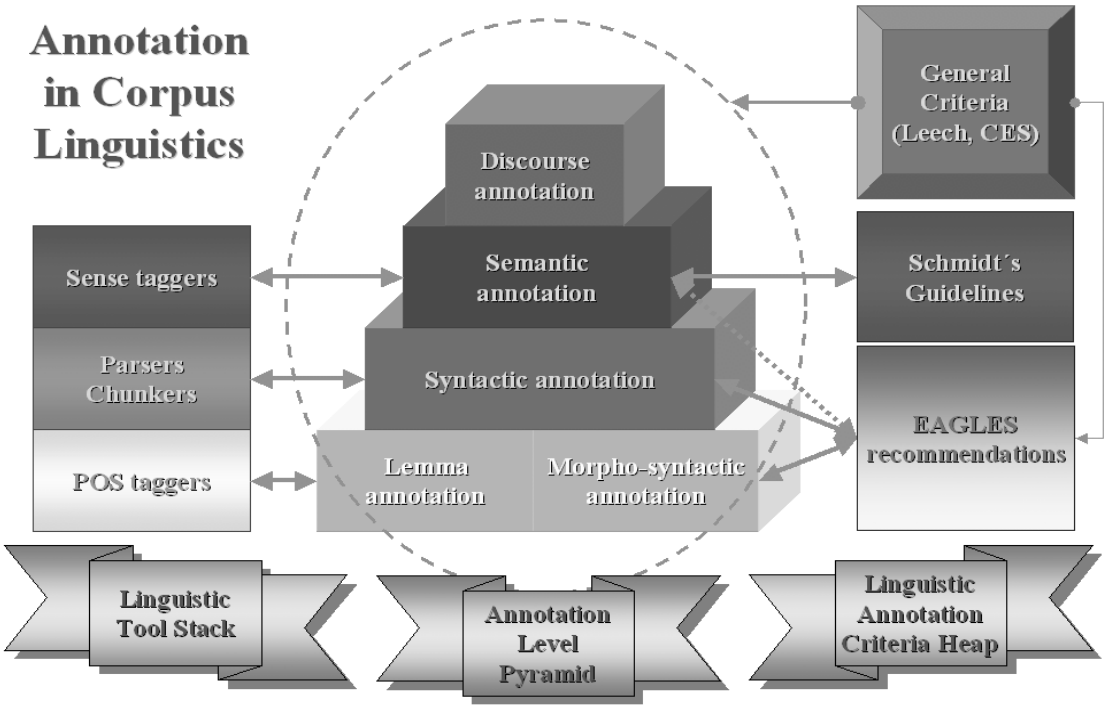

Figure 1: Annotation in Corpus Linguistics.

3 The ContentWeb Project

As mentioned before, the research here shown is being carried out within *ContentWeb*, a Spanish ministry MCyT funded project, which aims at the creation of an ontology-based platform that enables users to query e-commerce applications by using natural language, performing the automatic retrieval of information from web documents annotated with ontological and linguistic information. Besides, a prototype in the entertainment domain will be developed. *ContentWeb* objectives are sketched in Figure 2.

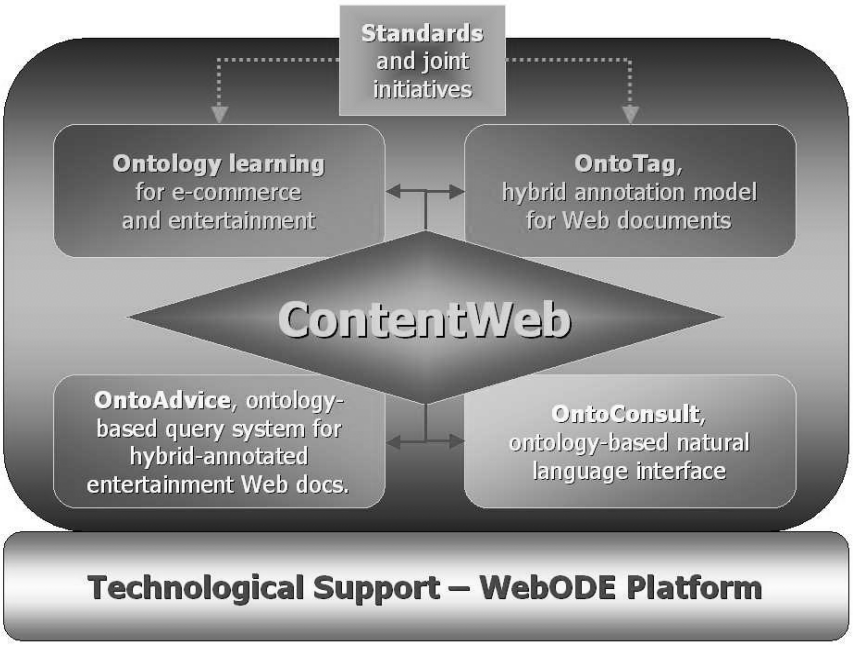

Figure 2: ContentWeb objectives.

4 OntoTag's Architecture

As shown in Figure 2, one of the objectives of *ContentWeb* is to develop *OntoTag*, a platform for the hybrid –linguistic and ontological– annotation of web documents. An *OntoTag* draft annotation example in RDF(S)/XML (also OWL compliant) can be found in Aguado–de Cea et al. (2002c). OntoTag's architecture proposal is presented in detail in Aguado–de Cea et al. (2003). Here, the overall architecture is shown in Figure 3 and each of its phases is briefly outlined.

Phase 1 – Cleaning: the textual information that a web document conveys is extracted and its markup information is stored away for the final re-construction of the page or for the use of this meta-textual information in the following steps of the annotation process.

Phase 2 – Annotation: the clean text produced in the cleaning phase is inputted to the different Spanish tools available in our laboratory, either licensed or online, i.e. FDG (Tapanainen & Järvinen, 1997).

Phase 3 – Decantation: the different kinds of annotation in the annotated text are split, according to the different linguistic description levels they belong to (lemma, POS, syntactic and semantic taggings and annotations[1]).

Phase 4 – Standardisation: the different decanted annotations are now mapped into a standard or guideline-compliant type of annotation: EAGLES (1996a; 1996b), CES (1999), MILE (2003), GDA (2002), etc.

Phase 5 – Zipping (or Level Merging): all the annotations that belong to the same level are united to bear a combined unique annotation per level.

Phase 6 – Merging (or Overall Hybrid Merging): this is a two-step phase: first, lemma, POS and syntactic annotations are *interweaved* and, then, in a second step, the *semantic merger* adds the hybrid (linguistic and ontologic) semantic level annotation, summing up all levels' annotations in one.

[1] In order to improve as much as possible the semantics made explicit by OntoTag´s annotations, other discourse or pragmatic level linguistic features are being explored at the moment, following a corpus-based approach and the theories issued in Álvarez-de-Mon (2003) and Mann & Thomson (1988) and, thus, not included still in the architecture.

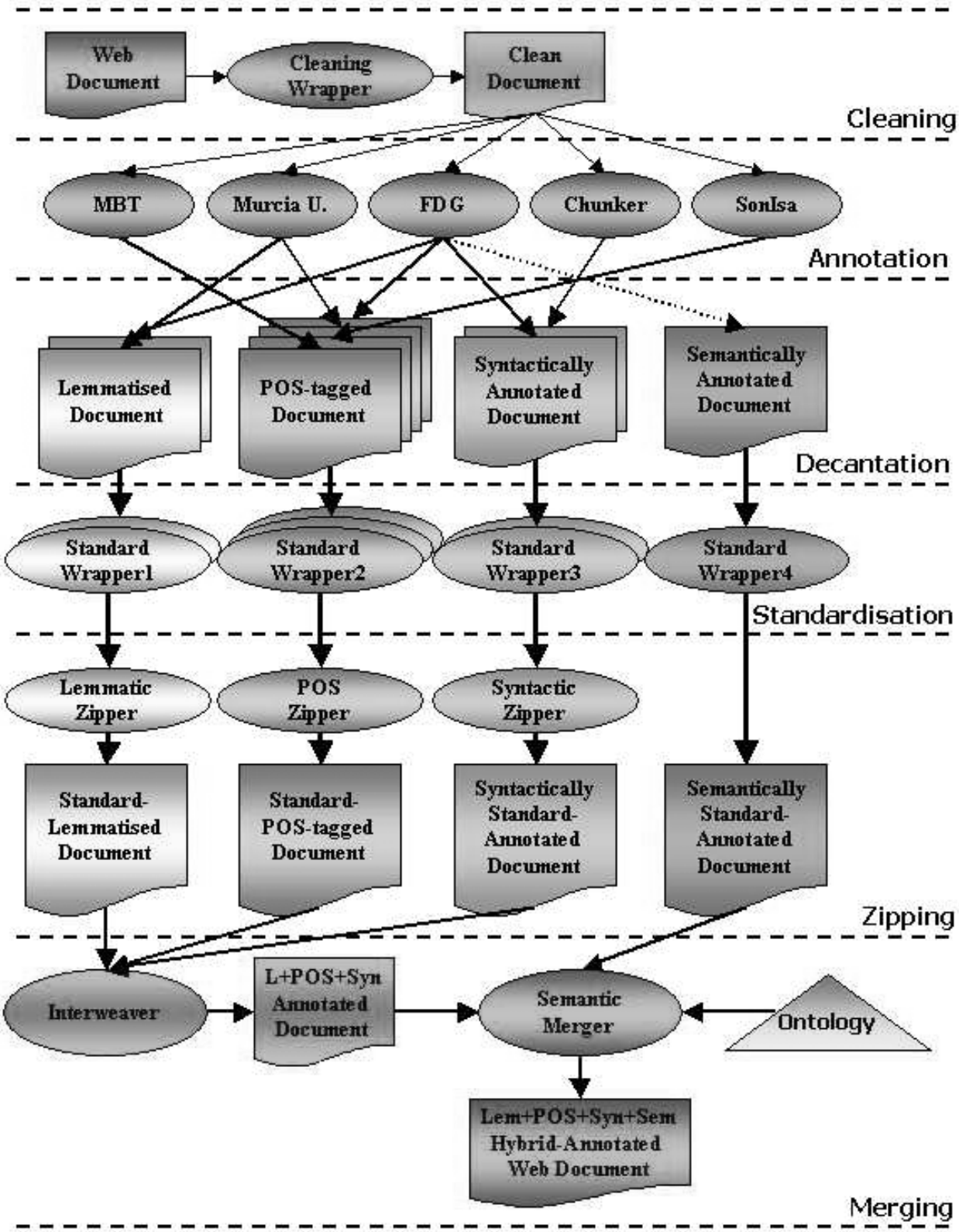

Figure 3: OntoTag's architecture proposal.

5 OntoTag's EAGLES & GDA Compliant Ontologies for Annotation

A set of ontologies (Gruber, 1993; Guarino & Giaretta, 1995) has been developed within WebODE (2003) to help OntoTag achieve the goal of standardisation, following the *EAGLES* (1996a, 1996b) *recommendations for the morpho-syntactic and syntactic annotation of corpora* and also the Global Document Annotation Initiative (GDA, 2002), for the aspects that where not handled or fixed by the EAGLES initiative.

According to EAGLES and OntoTag's multi-level philosophy, a linguistic level ontology has been created. Its taxonomical scheme is shown in Figure 4.

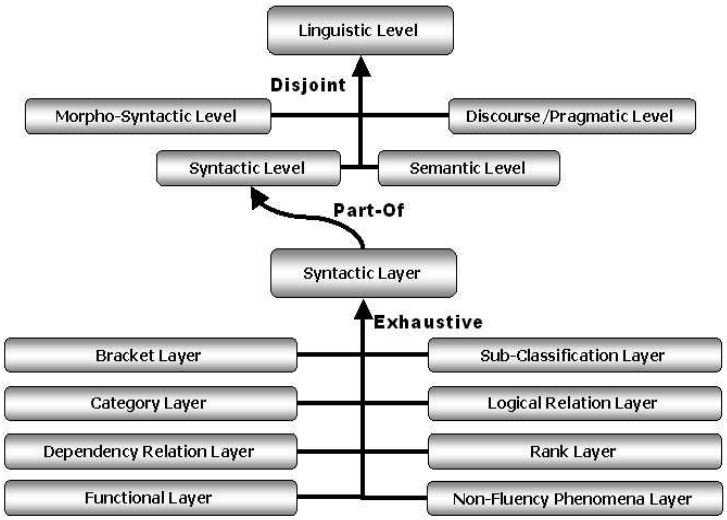

Figure 4: OntoTag's Linguistic Levels Ontology.

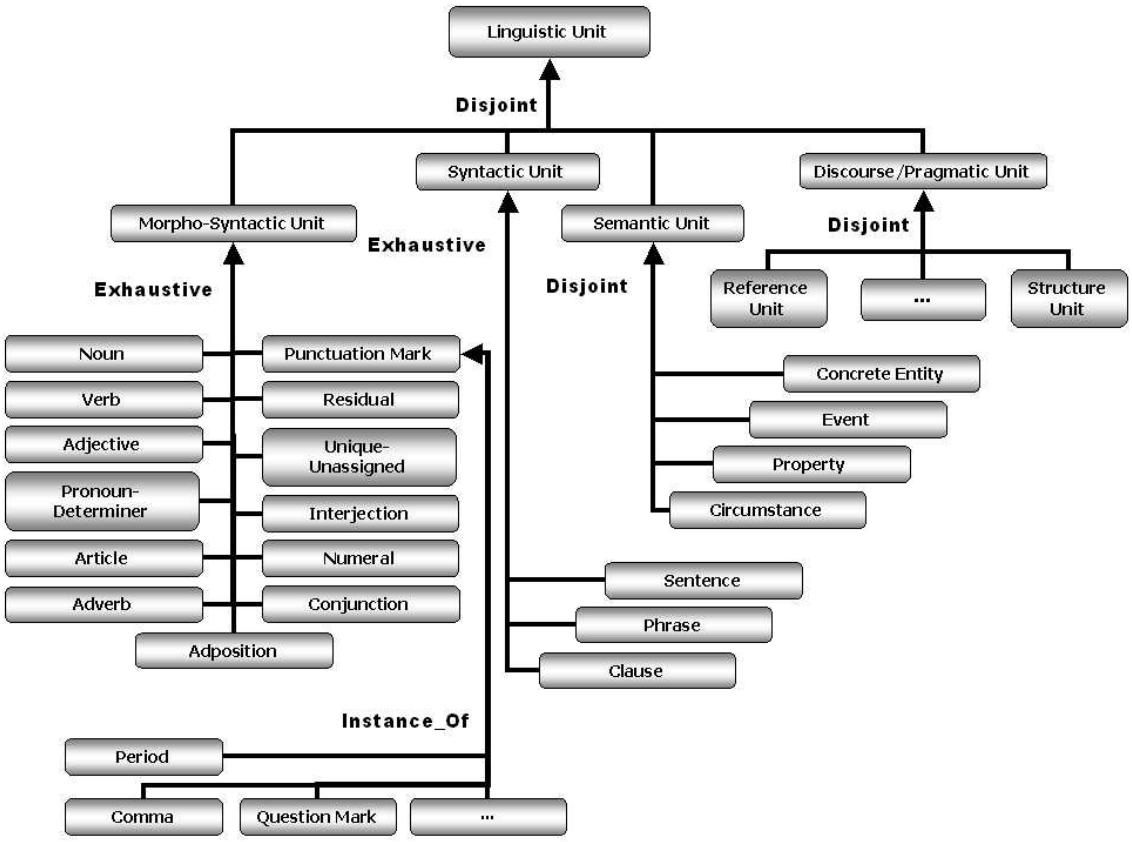

Figure 5: OntoTag´s Linguistic Unit Ontology (Partial Vew).

To comply with the EAGLES attribute-value formalism, two interconnected ontologies have been devised: one associated to linguistic units (whose first taxonomic level is shown in Figure 5) and another one related to linguistic attributes (partially shown in Figure 6).

1 **concrete_entity_linguistic_attribute**
1.1 **concrete_entity_morpho-syntactic_linguistic_attribute**
1.1.1 **{ gender, number, case, *np_function* }**
1.2 **concrete_entity_syntactic_linguistic_attribute**
1.2.1 **{ *syntactic_function* }**
1.3 **concrete_entity_semantic_linguistic_attribute**
1.3.1 **{ *thematic_role [GDA "Participant"]* }**
1.4 **concrete_entity_discourse/pragmatic_linguistic_attribute**
1.4.1 **{ person, possessive_number, *politeness* }**
2 **event_linguistic_attribute**
2.1 **event_morpho-syntactic_linguistic_attribute**
2.1.1 **{ finiteness, mood, tense, voice, status, *aspect, separability, reflexivity, auxiliary* }**
3 **property_linguistic_attribute**
3.1.1 **{ degree }**
4 **circumstance_linguistic_attribute**
4.1.1 **{ degree, *semantic_function [GDA "Spatiotemporal Relation" & "Other Semantic Relation"]* }**

Figure 6: Linguistic Attribute Ontology (Partial View)

Taxonomic and sub-categorisation attributes and values have been included in the linguistic unit ontology (LUO) instead of being in the linguistic attribute ontology (LAO), as deeper levels of specialization (subclassification) via *Is-A*, *Exhaustive* and *Disjoint* relationships (for example, two different subclasses appear in the LUO for Noun: Proper Noun and Common Noun, instead of an attribute type in the LAO). In both of them, most attribute values have been included via the *Instance-Of* relationship, instead of the *Is-A* relationship, if an instance cannot be distinguished from another when annotating a text (two different periods are always annotated with the same information, for example). The sets of this value instances have not been included here for the sake of brevity.

Anyway, these two ontologies (LAO and LUO) are interconnected by "ad-hoc" relationships such as, for example, *Has_Attribute(Linguistic Unit, Linguistic Attribute)*, instances of which are, for instance, *Has_Attribute(Noun, Number)* or *Has_Attribute (Verb, Tense)*. Besides, some constraints have been introduced by means of axioms such as:

- $\forall x$ *pronoun/determiner(x)* $\wedge$ *Value_of(x,subtype,partitive)* $\rightarrow$ *Value_of(x,type,determiner)*
- $\forall x$ *pronoun/determiner(x)* $\wedge$ *Value_of(x,subtype,personal/reflexive)* $\rightarrow$ *Value_of(x,type,pronoun)*

```
- <Term-Relation>
    <Name>Disjoint</Name>
    <Origin>Linguistic Attribute Group</Origin>
    <Destination>Linguistic Attribute</Destination>
    <Maximum-Cardinality>1</Maximum-Cardinality>
  </Term-Relation>
- <Group>
    <Name>Linguistic Attribute Group</Name>
    <Related-Concept>Property Linguistic
        Attribute</Related-Concept>
    <Related-Concept>Event Linguistic Attribute</Related-
        Concept>
    <Related-Concept>Concrete Entity Linguistic
        Attribute</Related-Concept>
    <Related-Concept>Circumstance Linguistic
        Attribute</Related-Concept>
  </Group>
- <Term-Relation>
    <Name>Subclass-of</Name>
    <Origin>Concrete Entity Morpho-Syntactic Linguistic
        Attribute</Origin>
    <Destination>Concrete Entity Linguistic
        Attribute</Destination>
    <Maximum-Cardinality>1</Maximum-Cardinality>
  </Term-Relation>
- <Term-Relation>
    <Name>Subclass-of</Name>
    <Origin>Concrete Entity Syntactic Linguistic
        Attribute</Origin>
    <Destination>Concrete Entity Linguistic
        Attribute</Destination>
    <Maximum-Cardinality>1</Maximum-Cardinality>
  </Term-Relation>
- <Term-Relation>
    <Name>Subclass-of</Name>
    <Origin>Concrete Entity Semantic Linguistic
        Attribute</Origin>
    <Destination>Concrete Entity Linguistic
        Attribute</Destination>
    <Maximum-Cardinality>1</Maximum-Cardinality>
  </Term-Relation>
- <Term-Relation>
    <Name>Subclass-of</Name>
    <Origin>Concrete Entity Discourse/Pragmatic
        Linguistic Attribute</Origin>
    <Destination>Concrete Entity Linguistic
        Attribute</Destination>
    <Maximum-Cardinality>1</Maximum-Cardinality>
  </Term-Relation>
```

Figure 7: LAO WebODE XML generated code.

Some other ontologies have been created to, i.e., represent the linguistic tools incorporated in OntoTag and their outputs, markup formats and Web content, and many "ad-hoc" relationships interconnect and complete OntoTag's underlying conceptual model; they are not included here for the sake of space. A portion of the WebODE

XML code generated from the ontologies[2] presented is shown in Figure 7.

6 Conclusions

This paper shows the results of the research carried out on finding an optimal model for the standardised semantic annotation of web pages (a *virtual corpus*); we have observed that joining together semantic annotation models from AI and the annotations proposed for every linguistic level from Corpus Linguistics is not only possible, but also needed and helpful to computers when facing the task of understanding the text contained in a document – a Semantic Web page.

The integration of these two fields (Corpus Linguistics and Artificial Intelligence) and their approaches entails many **advantages** for both of them. First of all, language and ontological resources (corpora, annotation tools, ontologies, etc.) will be more *reusable*, since both communities would profit from each other's advances, developments and results; *reuse of resources* will be improved, also, through the introduction of standardisation techniques, as OntoTag proposal illustrates. The second main advantage is that Corpus Linguistics researchers will be given for their work a virtual, freely available, annotated corpus[3] of (almost) infinite length: the Semantic Web. Involving both communities together (Corpus Linguistics and Artificial Intelligence) in the production of new models and schemes of annotation will entail a third benefit: the acceleration of the development of efficient annotation techniques (mainly by means of automation) or tools for corpus conversion, consistency checking and validation, for example, together with a wider and more consensual level of standardisation in this area.

However, **the main problem** for annotating web pages lies in the limitations imposed by current technologies; automatically obtaining compact, readable and verifiable pages is quite a hard task in itself, but its difficulty is even increased by the fact that it is neither fully specified nor delimited. The work being done in our laboratory is trying to bring some light upon this.

[2] Other export languages are generated by WebODE, such as DAML+OIL or Prolog and, soon, also OWL will be included.

[3] See Kilgarriff (2001), Gelbukh (2002).

Acknowledgements

The research described in this paper is supported by MCyT (Spanish Ministry of Science and Technology) under the project name: *"ContentWeb: Plataforma tecnológica para la Web Semántica: Ontologías, análisis de lenguaje natural y comercio electrónico"* – TIC2001-2745 ("ContentWeb: Semantic Web Technologic Platform: Ontologies, Natural Language Analysis and E-Business").

We would also like to thank the whole LIA for their help with the technological aspects of this paper.

References

Aguado-de Cea, G., Álvarez de Mon-Rego, I., Gómez-Pérez, A., Pareja-Lora, A. & Plaza-Arteche, R. 2002a. A Semantic Web Page Linguistic Annotation Model. *Semantic Web Meets Language Resources. Technical Report WS-02-16*. American Association for Artificial Intelligence. AAAI Press. Menlo Park, California, E.E.U.U.

Aguado-de Cea, G., Álvarez de Mon-Rego, I., Pareja-Lora, A. & Plaza-Arteche, R. 2002b. OntoTag: A Semantic Web Page Linguistic Annotation Model. *Proceedings of the ECAI 2002 Workshop on Semantic Authoring, Annotation and Knowledge Markup*. Lyon, Francia.

Aguado-de Cea, G., Álvarez de Mon-Rego, I., Pareja-Lora, A. & Plaza-Arteche, R. 2002c. RDF(S)/XML linguistic annotation of Semantic Web pages. *Proceedings of the 2nd Workshop on NLP and XML (NLPXML-2002)*. COLING'2002. Taipei, Taiwan.

Aguado-de Cea, G., Álvarez de Mon-Rego, I., Gómez-Pérez & A. Pareja-Lora, A. 2003. OntoTag: a Hybrid Platform for Meaningful Semantic Web Page Annotation. *Submitted to the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03).*

Álvarez de Mon-Rego, I. 2003. La cohesión del texto científico-técnico: un estudio contrastivo inglés-español. Universidad Complutense de Madrid (forthcoming).

Benjamins, V.R., Fensel, D., Decker, S., Gómez-Pérez, A. 1999. (KA)^2: Building Ontologies for the Internet: a Mid Term Report. *IJHCS,*

International Journal of Human Computer Studies, 51, pp. 687–712.

Berners-Lee, T., Fischetti, M. 1999. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*. Harper. San Francisco.

Buitelaar, P., Bryant, B., Ide, N., Lin, J., Pareja-Lora, A., Wilcock, G. 2003. The Roles of Natural Language and XML in the Semantic Web. *Language and Linguistics* (forthcoming).

CES. 1999. *Corpus Encoding Standard*. http://www.cs.vassar.edu/CES/

EAGLES. 1996a. *EAGLES: Recommendations for the Morphosyntactic Annotation of Corpora*. EAGLES Document EAG--TCWG—MAC/R.

EAGLES. 1996b. *EAGLES: Recommendations for the Syntactic Annotation of Corpora*. EAGLES Document EAG--TCWG—SASG/1.8.

EsperOnto. 2003. *Esperonto Services IST-2001-34373 Deliverable on Annotation*. http://www.esperonto.net/ (forthcoming).

GDA. 2002. Global Document Annotation Initiative: The GDA Tag Set. http://www.i-content.org/GDA/tagset.html

Gelbukh, A., Sidorov, G., Chanona-Hernández, L. 2002. Corpus virtual, virtual: Un diccionario grande de contextos de palabras españolas compilado a través de Internet. Gonzalo, J., Peñas, A., Ferrández, A. (eds.) *Proceedings of the IBERAMIA 2002 Workshop on Multilingual Information Access and Natural Language Processing*. Sevilla, Spain. pp: 7–13

Gruber, R. 1993. A translation approach to portable ontology specification. *Knowledge Acquisition*, 5, pp. 199–220.

Guarino, N., Giaretta, P. 1995. Ontologies and Knowledge Bases: Towards a Terminological Clarification. *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, N. Mars, ed., IOS Press, Amsterdam, pp. 25–32.

ISLE. 2003. http://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm

Kilgarriff, A. 1998. *SENSEVAL:* An Exercise in Evaluating Word Sense Disambiguation Programs. *Proceedings of LREC*, Granada, Spain, pp. 581–588.

Kilgarriff, A. & Rosenzweig, J. 2000. English SENSEVAL: Report and Results. *Proceedings of LREC*. Athens, Greece.

Kilgarriff, A. 2001. Web as Corpus. http://www.itri.bton.ac.uk/~Adam.Kilgarriff/PAPERS/corpling.txt

Leech, G. 1997a. Introducing corpus annotation. *Corpus Annotation: Linguistic Information from Computer Text Corpora*, R. Garside, G. Leech & A. M. McEnery, ed., Longman, London.

Leech, G. 1997b. Grammatical tagging. Garside R., Leech, G., McEnery, A. M. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

Luke S., Heflin J. 2000. *SHOE 1.01. Proposed Specification*. SHOE Project. http://www.cs.umd.edu/projects/plus/SHOE/spec1.01.htm

Mann, W & Thomson, S. 1988. *Rhetorical Structure Theory: Toward a functional theory of text organization*. Text Vol.18, 3, pp. 243–281.

McEnery, A. M., Wilson, A. 2001. *Corpus Linguistics: An Introduction.* Edinburgh University Press, Edinburgh.

Motta, E., Buckingham Shum, S. Domingue, J. 1999. Case Studies in Ontology-Driven Document Enrichment. *Proceedings of the 12th Banff Knowledge Acquisition Workshop*, Banff, Alberta, Canada.

Nirenburg, S. and Raskin, V. 2001. *Ontological Semantics (Draft)* Visited on September 2001. http://crl.nmsu.edu/Staff.pages/Technical/sergei/book/index-book.html

Schmidt, K. M. 1988. Der Beitrag der begriffsorientierten Lexicographie zur systematischen Erfassung von Sprachwandel und das Begriffswörterbuch zur mhd. Epik. *Mittelhochdeutsches Wörterbuch in der Diskussion*, ed. by Bachofer, W. Tübingen: Max Niemeyer, 35–49.

SHOE. 2002. http://www.cs.umd.edu/projects/plus/SHOE/KnowledgeAnnotator.html

Staab, S., Angele, J., Decker, S., Erdmann, M., Hotho, A., Mädche, A., Schnurr, H.-P., Studer, R. 2000. Semantic Community Web Portals. *WWW'9*. Amsterdam.

Tapanainen, P., Järvinen, T. 1997. A non-projective dependency parser. *Proceedings of the 5th conference on Applied Natural Language Processing*. Washington D.C.: Association for Computational Linguistics, 64–75.

WebODE. 2003. http://delicias.dia.fi.upm.es/webODE/

Wilson, A., Thomas, J. 1997. Semantic Annotation. *Corpus Annotation: Linguistic Information from Computer Text Corpora*, R. Garside, G. Leech & A. M. McEnery, ed., Longman, London.